List of Implemented and Ongoing Projects

Part 1: Implemented Projects

SABA Sports

Project 1: Real-time Sports Betting Fraud Detection and Prediction System

Project Goal: To build a large-scale Machine Learning and Deep Learning system capable of analyzing user behavior and real-time match statistics. The objective is to predict and flag suspicious activities that could be fraudulent or involve match-fixing, with the capacity to handle data from millions of concurrent users during live sports events.

Role and Responsibilities (Data Architect & Lead Data Scientist): In this project, I held a dual role:

- **Data Architect (Phase 1):** Independently responsible for surveying, designing, and re-implementing the entire on-premise data architecture, laying the foundation for large-scale analysis and model development.
- **Technical Lead (Phase 2):** Led a team of 4 members, with primary responsibility for researching, designing algorithms, and deploying Machine Learning/Deep Learning models into the production environment.

Methodology and Solution Architecture:
The project was executed in two strategic phases:
Phase 1: Designing and Building a High-Performance Data Platform (On-Premise):

- **Challenge:** The existing system could not handle the load requirements of millions of users and lacked a flexible structure for complex data analysis.
- **Architectural Solution:**
    - **Datalake:** I proposed and implemented a Datalake architecture using Apache Spark. The focus was on designing an optimized Landing Zone for high-speed data ingestion, while being flexible enough to integrate preprocessing workflows and apply ML/DL models directly on raw data.
    - **Data Warehouse:** Designed the initial Data Warehouse on PostgreSQL with a Star Schema model, prioritizing query speed for existing business reports.
    - **Data Orchestration:** All data processes (ETL/ELT) were automated and orchestrated using Apache Airflow, ensuring stability and reliability.

Phase 2: Developing and Deploying the Fraud Prediction Model:

- **Problem:** A model was needed to detect anomalous betting patterns (outliers)

from user data, which is a critical indicator of match-fixing.

- **Model Solution:**
  - As the team lead, I proposed a hybrid approach combining a Tensor-based model for multi-dimensional interaction analysis and XGBoost. This approach helped capture complex, non-linear relationships in user behavior while ensuring fast processing and high accuracy for early warnings.
- **Data Warehouse Restructuring:**
  - As the model generated new predictive data dimensions, I realized the initial Star Schema would not be flexible enough. I proposed and led the transition of the Data Warehouse architecture to a Snowflake Schema. This change made the system easily scalable and allowed for the integration of model prediction results into reports without disrupting the existing structure.

Achievements:

- **High Accuracy:** The model achieved up to **78%** accuracy in identifying suspicious events and potential match-fixing risks.
- **Real-time Performance:** The system was optimized to achieve near real-time prediction performance, with latency ranging from only **0.55s to 0.98s** per request.
- **Superior Scalability:** The architecture was successfully designed to handle data from **20 matches** simultaneously and serve millions of users during peak hours.
- **Effective Fraud Detection:** The system successfully identified over **65%** of the malicious attack behaviors found in the validation datasets.

Technologies and Models Used:

- **Languages:** Python, SQL
- **Data Processing & Orchestration:** Apache Spark, Apache Airflow, Hive
- **Database:** PostgreSQL
- **ML/DL & Algorithms:** XGBoost, PyTorch, LSTM, Tensor-based Models, Outlier Detection
- **Architecture:** Datalake, Data Warehouse (Star & Snowflake Schema), ETL/ELT, Real-time Data Pipelines

Project 2: Formation Tactics Evaluation and Recommendation System using a Recommendation System (Solo Project) (SABA Sport)

Project Goal: To develop an in-depth analysis system capable of quantifying and evaluating the effectiveness of different team formation tactics (home vs. away). The main objective is to use Machine Learning techniques to predict the compatibility and effectiveness of one formation against another, thereby providing valuable tactical

suggestions before and during the match.

Role and Responsibilities (Data Scientist & ML Engineer - Independent): As the sole implementer of this project, I was responsible for:

- **Ideation and Research:** Developed the core idea of applying a hybrid model of XGBoost and Matrix Factorization to the problem of tactical evaluation.
- **Data Model Design:** Built the data structure to "model" formations and tactics. Each formation was treated as an "item," and the head-to-head result/performance was treated as a "rating."
- **Model Building and Training:** Implemented algorithms to learn from historical head-to-head data and optimized them.
- **Testing and Optimization:** Evaluated the accuracy of tactical recommendations based on actual match results and optimized the system for high performance.

Methodology and Solution Architecture:

- **High-Performance ETL for Data from OPTA:**
  - The system uses highly accurate and in-depth player data from OPTA.
  - To meet the real-time analysis requirement, I designed a high-performance ETL (Extract, Transform, Load) process capable of processing, cleaning, and structuring data from 10 simultaneous matches.
- **Hybrid Core Model:**
  - The heart of the system is a hybrid model architecture that combines two powerful techniques for a comprehensive view:
    - **XGBoost for Player Action Evaluation:** An XGBoost model was applied to analyze and evaluate the effectiveness of individual player actions in real-time (e.g., a pass, a tackle, a shot). This model returns an "effectiveness" score for each action.
    - **Matrix Factorization for Formation & Interaction Evaluation:** The Matrix Factorization technique was used to analyze the interaction matrix between players on the field. The goal was to discover latent factors representing the synergy between players and between players-and-tactics. This helps to quantitatively assess how well players coordinate within a specific formation.
- **Result Synthesis:**
  - Insights from both models were combined to provide a comprehensive assessment: XGBoost shows "who is doing what well," while Matrix Factorization shows "who is coordinating well with whom." This combination creates a detailed picture of the team's overall tactical effectiveness.

Achievements:

- **Accurate Tactical Recommendations:** The system achieved an accuracy of **over 75%** in predicting the formation with a greater advantage based on past head-to-head data.
- **Instant Analysis:** Provided evaluation and comparison between tactical pairs with near real-time performance, with latency of only **0.55s - 0.98s**.
- **Counter-Tactic Detection:** The system is capable of identifying "counter" or "mismatched" formation pairings that might be difficult to spot through conventional analysis.

Technologies and Models Used:

- **Languages & Frameworks:** Python, PyTorch
- **Models/Algorithms:** XGBoost, Recommendation Systems (Matrix Factorization), Poisson Distribution

DXC Technology

Project 3: Real-time Optimized Customer Information Storage System (DXC Technology)

Project Goal: To build a robust and scalable data platform dedicated to processing and updating sensitive customer information (e.g., health status, pre-existing conditions) in near real-time. The main objective is to ensure data integrity and continuous updates for millions of users, while providing a solid foundation for automated reporting and in-depth data analysis.

Role and Responsibilities (Data Architect & Senior Data Engineer): As the lead designer and implementer, I was responsible for:

- **End-to-End Architecture Design:** Conceptualized and designed the entire architecture from the data ingestion stream to the storage & serving layer.
- **Building Data Pipelines:** Directly designed and implemented high-performance data pipelines (ETL/ELT), ensuring the ability to handle high-frequency record updates.
- **Storage Architecture:** Decided on and built a hybrid Data Warehouse / Lakehouse storage architecture, optimized for both structured reporting and flexible analysis.
- **Data Modeling:** Designed data models within the Warehouse/Lakehouse to ensure query performance and data consistency.
- **Performance Optimization:** Focused on tuning the system to achieve the highest possible throughput, meeting the requirement for second-level updates

across millions of records.

Methodology and Solution Architecture:

- **Real-time Ingestion Pipelines:**
  - The core of the project was designing ETL/ELT pipelines capable of handling continuous streams of updated data. This architecture was optimized to minimize latency, allowing sensitive customer information to be reflected on the system almost immediately after a change occurs.
- **Hybrid Storage Architecture:**
  - I proposed and implemented an advanced Data Warehouse / Lakehouse architecture.
  - The Data Lake layer was used to store raw, flexible data for complex analysis and future Machine Learning model training.
  - The Data Warehouse layer was built on top, containing structured, cleaned, and aggregated data. This layer was specifically optimized for generating automated reports and allowing business analysts to easily query and identify special customer cases or records.
- **Scalability & High-Throughput:**
  - The entire system was designed with a "scalability first" philosophy, ready to scale to millions of users without performance degradation.
  - Parallel and distributed data processing techniques were applied to ensure the system could handle a large volume of concurrent updates, achieving throughput measured in seconds.

Achievements:

- Successfully implemented an effective **Data Warehouse / Lakehouse** solution, which became the foundation for automated reporting and allowed for the rapid detection of customer profiles requiring special attention.
- Designed a system that achieved **extremely high processing throughput**, ensuring data could be updated in near real-time (measured in seconds) across millions of concurrent customer records.

Technologies and Models Used:

- ETL/ELT, Real-time Data Pipelines, Data Warehousing, Lakehouse Architecture, Data Modeling, High-Volume Data Processing.

Project 4: Modernization and Migration of ETL System from On-Premise (Informatica) to Azure Data Factory (ADF)

Project Goal: To modernize the client's entire data processing infrastructure by migrating ETL (Extract, Transform, Load) processes from an outdated On-Premise

system (Informatica) to the Azure Data Factory (ADF) cloud platform. The goal was not just to migrate the system but also to enhance scalability, improve reliability, and reduce operational costs for the client.

Role and Responsibilities (Senior Data Engineer & Data Migration Specialist): As the lead technical person for the migration process, my role included:

- **Analysis and Reverse-Engineering:** Primarily responsible for analyzing existing ETL processes on Informatica. Especially for legacy processes with no documentation, I had to independently understand and reverse-engineer the logic to ensure no business requirements were missed.
- **Redesign on the Cloud:** Redesigned and re-implemented the entire business logic, optimizing it using the native components and features of Azure Data Factory.
- **System Documentation:** Built a detailed and comprehensive technical documentation set from scratch for the new ETL system on the cloud, to serve future operation and maintenance.
- **Testing and Quality Assurance:** Performed parallel testing and data reconciliation processes to ensure the output of the new ADF system perfectly matched the old system.

Methodology and Solution Architecture:
The migration process was carried out in 3 main phases:

1. **Discovery & Analysis Phase:**
    - Conducted an audit of all workflows and mappings on the existing Informatica PowerCenter system.
    - Identified data sources, destinations, complex transformations, and dependencies between processes.
    - The main challenge was focusing on "reverse-engineering" and documenting the logic of critical but undocumented ETL processes.
2. **Design & Re-implementation on Azure Phase:**
    - Instead of a "lift-and-shift" 1-1 copy, I applied a re-architecting approach.
    - Each logical process from Informatica was redesigned to fully leverage the power of Azure Data Factory, for instance, using Data Flows for complex transformations and integrating with other Azure services like Azure Blob Storage and Azure SQL Database.
3. **Testing & Handover Phase:**
    - After deployment, the pipelines on ADF were run in parallel with the jobs on Informatica.
    - The output data was carefully reconciled to ensure 100% integrity and accuracy.

    - After confirmation, the new system was handed over to the client along with the detailed logical documentation that had been created.

Achievements:

- Completed 100% migration of the client's ETL processes from Informatica to Azure Data Factory on schedule.
- Successfully built and delivered a comprehensive technical documentation set, helping the client to easily understand, operate, and maintain the new system.
- Successfully modernized the client's data infrastructure, helping them increase flexible scalability, improve process reliability, and significantly reduce costs associated with maintaining on-premise servers.

Technologies and Models Used:

- **Cloud Platform:** Microsoft Azure
- **Data Integration Services:** Azure Data Factory (ADF), Informatica PowerCenter
- **Concepts:** ETL/ELT, Data Migration, Cloud Modernization, Reverse-Engineering, Data Warehousing.

Kimberly-Clark Viet Nam

Project 5: Market Analysis and Forecasting to Build Product Strategy on E-commerce Platforms (Kimberly-Clark)

Project Goal: To build a predictive analytics system aimed at providing a competitive edge on domestic and international e-commerce platforms. The main objective is to apply Machine Learning and Deep Learning models to analyze market data, forecast products with breakthrough potential, and provide strategic reports to guide decisions on product promotion and development.

Role and Responsibilities (Business Analyst & Data Scientist): In this role, I was primarily responsible for:

- **Forecasting Model Development:** Directly developed and applied forecasting models using Machine Learning and Deep Learning techniques to analyze market data.
- **Competitive Analysis:** Regularly conducted important analytical reports, comparing the performance of internal products with competitors in the e-commerce market.
- **Strategic Report Creation:** Responsible for creating periodic and ad-hoc reports to inform strategies and promotional campaigns for products on various e-commerce channels.
- **Data-Driven Consulting:** Provided data-based recommendations to improve the

accuracy of forecasts and support strategic planning for the organization.
- **Data Mining:** Applied AI and data mining methods to extract key insights into e-commerce market dynamics and product performance.

Methodology and Solution Architecture:

- **Data Collection and Integration:** The system aggregates data from various sources, including internal sales data, market data from e-commerce platforms, and information on competitor activities.
- **Building a Product Potential Forecasting Model:**
  - Used Machine Learning models to analyze current trends and historical data.
  - The goal of these models is to identify the characteristics or patterns of successful products, thereby forecasting new products with the potential to make a breakthrough in the market.
- **Performance and Competitive Analysis:**
  - Established an automated reporting process using BI tools.
  - These reports provide a multi-dimensional view, comparing key performance indicators (KPIs) of the company's products with major competitors, helping to identify strengths, weaknesses, and opportunities.
- **Providing Insights for Marketing Activities:**
  - The predictive insights from the models were translated into specific reports and recommendations.
  - These reports serve as crucial input, supporting the marketing department in designing and implementing effective product promotion campaigns on e-commerce channels.

Key Results:

- Provided predictive insights into market trends and product potential.
- Delivered useful and actionable competitive analysis reports.
- Supported data-driven decisions for product promotion activities on e-commerce platforms.

Technologies and Models Used:

- **BI Reporting Tools**
- **E-commerce Analytics**
- **Models/Algorithms:** XGBoost, Poisson Distribution, Linear Regression, Logistic Regression

Part 2: Ongoing Projects

Project: SABA InsightAI - Automated Match Context Analysis and Interpretation

System using LLM

Project Goal: To create a new premium analytical product for SABA Sports' B2B clients. This system will not only provide raw data but will use an LLM to synthesize, interpret, and generate narratives about the context of an ongoing match in real-time. The goal is to provide clients (bookmakers, sports media companies) with deep, easy-to-understand, and immediately actionable insights, helping them make more effective risk management and end-user engagement decisions.

Problem to Solve: SABA's clients receive a huge amount of data (odds, match statistics, user behavior). However, they have to expend human resources (analysts) to connect these data points, understand the "story" behind the numbers, and react in a timely manner. This process is slow and can miss important signals.

Role and Responsibilities (yours): As the Lead Data Scientist for this project, I'm responsible for:

- **End-to-End Architecture Design:** Build the overall architecture, from integrating SABA's existing real-time data streams into the LLM processing system.
- **Model Selection and Fine-tuning:** Select a foundational LLM (e.g., Llama 3, Gemini, Mistral) and fine-tune it on SABA's historical data repository (match data, news, sports commentary) so the model "understands" the sports domain deeply.
- **RAG (Retrieval-Augmented Generation) Architecture Development:** Build a RAG system to provide the LLM with accurate and up-to-date contextual information in real-time, preventing model hallucination.
- **Prompt Engineering:** Design complex prompt chains to instruct the LLM to act as a sports analysis expert, capable of synthesizing multiple information sources and providing in-depth commentary.
- **Integration and Optimization:** Work with the engineering team to integrate the solution into SABA's existing products and optimize for the lowest possible latency.

Methodology and Solution Architecture:
The system will operate in 4 layers:

1. **Multi-Source Ingestion Layer:**
    - **Structured Data:** Leverage SABA's existing real-time data pipelines, capable of handling up to 20 matches simultaneously. Data includes: fluctuating odds, match statistics (shots, cards, possession), and suspicious user behavior data.
    - **Unstructured Data:** Build new crawlers/scrapers to fetch data from external sources: sports news, social media posts (Twitter/X), press releases from

clubs, live commentary from major news sites.

2. **Processing & Embedding Layer:**
    - All data (both structured and unstructured) will be passed through a processing pipeline and converted into vector embeddings using models like text-embedding-ada-002 or other open-source alternatives.
    - These vectors are stored in a Vector Database (Chroma), allowing for extremely fast semantic similarity searches.
3. **Intelligent Interpretation Layer - The RAG + LLM Core:**
    - When a significant event occurs in a match (e.g., a red card, a surprise goal, a sudden large bet on an unusual market), the RAG system is triggered.
    - **Retrieval:** The system will query the vector database to find all information related to that event (e.g., does the player who just received a red card have a history of foul play? Any recent news about his injuries? What is the social media reaction to the referee's decision? Is there anything unusual about the odds history for this match?)
    - **Generation:** The retrieved information is packaged into a complex prompt and sent to the fine-tuned LLM. The LLM will be tasked: "You are a sports risk analyst. Based on the following data, write a summary (2-3 sentences) explaining the significance and potential risks of the event that just occurred."
4. **Output & Interaction Layer:**
    - **Automated Alerts:** Instead of a blunt alert like "78% risk of manipulation," the system will send a context-rich alert: "Warning: A spike in bets on Over 3.5 goals coincides with Team A's key defender receiving a red card. Social media analysis shows discontent with the referee's decision. Recommend reviewing the risk for this market."
    - **Interactive Q&A Interface:** SABA's clients can open a chat interface and ask questions in natural language: "Summarize the unusual events in the match between X and Y," "How is player Z's form in this match?", "Compare the odds of this match with the last 5 matches of team X."

Progress: Testing on Product

Conclusion

The completed and ongoing projects not only contribute to personal achievements but also bring value to the community and the professional field. Continuing to maintain and develop the ongoing projects will help achieve more positive results in the future.